A New Fast Parallel Statistical Measurement Technique for Computational Cosmology

Robert J. Thacker*
Department of Physics, Engineering Physics and Astronomy,
Queen's University, Kingston, Ontario, Canada K7L 3N6
E-mail: thacker@astro.queensu.ca
*Corresponding author

H. M. P. Couchman
Department of Physics and Astronomy
McMaster University
1280 Main St. West, Hamilton, Ontario, Canada L8S 4M1
E-mail: couchman@physics.mcmaster.ca

Abstract: Higher order cumulants of point processes, such as skew and kurtosis, require significant computational effort to calculate. The traditional counts-in-cells method implicitly requires a large amount of computation since, for each sampling sphere, a count of particles is necessary. Although alternative methods based on tree algorithms can reduce execution time considerably, such methods still suffer from shot noise when measuring moments on low amplitude signals. We present a novel method for calculating higher order moments that is based upon first top-hat filtering the point process data on to a grid. After correcting for the smoothing process, we are able to sample this grid using an interpolation technique to calculate the statistics of interest. The filtering technique also suppresses noise and allows us to calculate skew and kurtosis when the point process is highly homogeneous. The algorithm can be implemented efficiently in a shared memory parallel environment provided a data-local random sampling technique is used. The local sampling technique allows us to obtain close to optimal speed-up for the sampling process on the Alphaserver GS320 NUMA architecture.

Keywords: Cosmology; statistics; parallel programming; simulation

Reference to this paper should be made as follows: Thacker, R.J. and Couchman, H.M.P. (2006) 'A New Fast Parallel Statistical Measurement Technique for Computational Cosmology' Int. J. High Performance Computing and Networking, Vol. 3, Nos. 2, pp.xx–yy.

Biographical notes: Robert Thacker is an Adjunct Assistant Professor of Physics at Queen's University, Kingston, Ontario, Canada. He received his PhD degree from the University of Alberta in 1999. Since 1995 he has been conducting research in cosmology and is a specialist in numerical simulation and parallel computing.

Hugh M. P. Couchman is a Professor of Physics at McMaster University, Hamilton, Ontario, Canada. He received his PhD from Cambridge University in 1986. Since that time he has pursued research in computational cosmology and is co-developer of a number of tools that have aided breakthroughs in numerical modelling. Since 2004 he has served as the Scientific Director of the SHARCNET computing collaboration and is recognized as a pioneer of a number of Canadian HPC initiatives.

1 Introduction

Modern cosmology, the study of the large scale structure and evolution of our universe (Peacock, 2000), has advanced to the point where we can now answer some very fundamental questions about the distribution of matter within our universe. Ever since Einstein postulated the theory of General Relativity and, together with De Sitter (Pais, 1982), showed how it could be applied to the universe as a whole, generations of physicists have pondered on the question of what is the overall geometry of our universe. Within the past few years observations of the relic microwave radiation from the "Big Bang" (Bennett et al., 2003) have shown that the universe exhibits a geometry quite unlike that expected from theoretical prejudices alone.

Although on the largest scales the distribution of matter within our universe is both homogeneous and isotropic, on smaller scales—less than 1/20th the size of our visible universe—it is highly inhomogeneous. Even though the matter distribution of the universe was exceptionally smooth 300,000 years after the creation event (Kolb and Turner, 1990), over billions of years the ubiquitous attraction of the gravitational force amplifies the minute fluctuations in the early matter distribution into the structure we see today. Moreover, the current best theories of structure formation suggest that the matter distribution we observe is formed in a 'hierarchical clustering' manner with the small structures merging to form larger ones and so forth (Peacock, 2000). This growth of structure is accelerated by an unseen massive 'dark matter' component in our universe. Although dark matter cannot be observed directly, there is sufficient evidence within observations to conclusively infer its existence. Modifications to Newton's equations, to change gravitational accelerations on large scales, have had limited success, and cannot presently be cast in a form compatible with General Relativity (Sanders and McGaugh, 2002).

Understanding the distribution of matter within our local universe can tell us much about the cosmic structure formation process. While on the very largest scales gravity is the dominant force, on smaller scales gas pressure forces, from the gaseous inter-galactic (IGM) and interstellar mediums (ISM), can play a significant role. In clusters of galaxies, for example, hydrodynamic forces produced by the IGM lead to a distribution of gas that is held close to hydrostatic equilibrium. Indeed, understanding the interaction between the ISM and the stars that condense out of it, is currently one of the hottest research areas in cosmology (Thacker and Couchman, 2001). Since if we can understand this process we are much closer to being able to infer how the galaxies we observe relate to the underlying distribution of dark matter that dominates the evolution of structure.

Although we are yet to absolutely determine the relation between galaxies and dark matter, measuring the distribution of galaxies is the only way of infering the distribution of all matter (visible or not). Measurements of the speed of recession of local galaxies, led Hubble (1929) to form the distance-redshift relation now know as 'Hubble's Law', which has become a bedrock for the development of cosmological theory. Although modern surveys of galaxies use an updated, and more accurate, form of the distance-redshift relation to uncover the spatial distribution of galaxies, the principles involved remain the same as those used by Hubble.

Aided by highly automated observing and computer driven data analysis, a new generation of high quality galaxy redshift surveys is mapping our local Universe with exquisite precision. The 2 degree field (2df, 2004) and Sloan Digital Sky Survey (SDSS, 2004) provide astronomers with a survey of the local universe out to a redshift of $z \simeq 0.3$, and contain over 200,000 and one million (when complete) redshifts respectively. In figure 1 we show the distribution of galaxies for the 2dF survey to give an visual impression of the type of inhomogeneity observed.

Traditionally, one of the primary goals of analysis of redshift surveys is the calculation of the two point autocorrelation function (2-pt CF). The large sample volumes provided by 2dF and the SDSS have allowed the 2-pt CF to be calculated with great accuracy. While the initial conditions produced by the "Big Bang" are widely believed to exhibit Gaussian statistics (*e.g.* Kolb and Turner, 1990), the formation of structure by gravitational instability introduces non-Gaussian features into the statistics of the matter distribution. Hence, the 2-pt CF cannot be a complete descriptor of the underlying matter distribution at late times. Astronomers were aware of this issue comparatively early in the development of the field, and the theoretical basis for calculating higher order statistics was developed through the 1970's (see Peebles (1980) for a detailed summary).

Early attempts to measure higher order moments of the mass distribution, via the counts-in-cells method (again see Peebles (1980)), suffered from inadequate sample size. Because higher order moments tend to be progressively dominated by the most dense regions in a given sample, ensuring that adequate sampling has been performed is of utmost importance. Ensuring low sample variance is also necessary, and given one sample the only way to check this is to analyse sub-samples, which rapidly depletes the available information.

From a theoretical perspective, higher order statistics are interesting in relation to gravitational perturbation theory and the evolution of non-linear gravitational clustering. Analyses examining the accuracy of numerical simulation methods often rely upon higher order statistics. This is especially important in the study of gravitational clustering in 'scale free' universes (Couchman and Peebles, 1998). The development of fast, parallel, statistical algorithms is vital to progress in this arena. While the development of parallel simulation algorithms has advanced forward rapidly (*e.g.* Thacker et al., 2003) development of parallel analysis tools has lagged behind. This is partially due to the fact that the benefits of developing a parallel

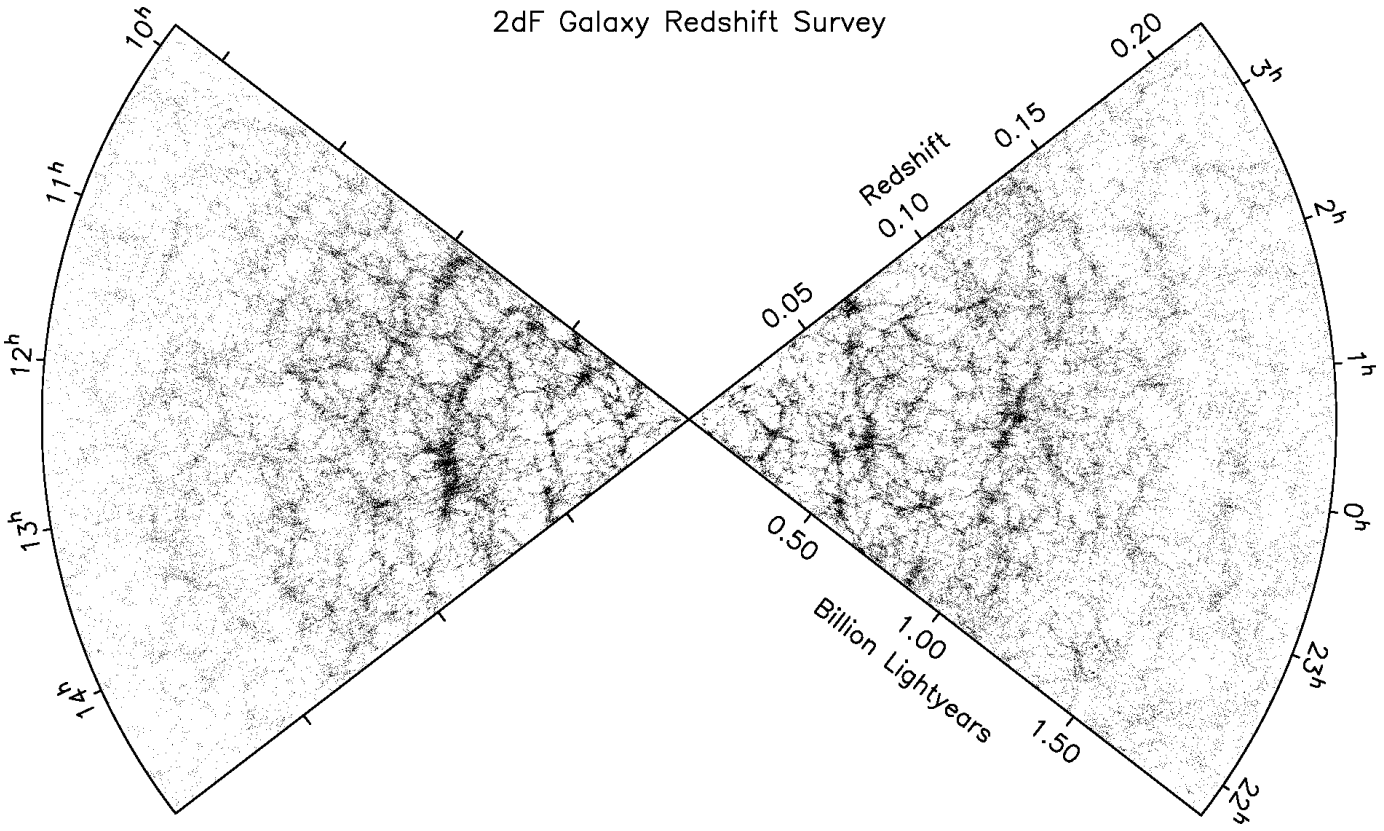

Figure 1: Distribution of galaxies in the two main slices from the 2dF galaxy redshift survey. Each point represents a galaxy, and they combine to trace filament and wall structures in three dimensions. The geometry of the distribution is directly related to the statistical properties of the conditions in our universe following the "Big Bang".

analysis code can be shorted lived because the required analyses can change rapidly (much faster than the simulation algorithms themselves). The rapid development times available on shared memory parallel machines make them an ideal complement to large distributed memory machines which most simulations are now run on.

Although throughout this paper we discuss the application of our new method to cosmology, it can be applied equally well to the statistics of any point process. Indeed the terms 'particle' and 'point' are often used interchangeably. The method can also be modified to apply to different dimensions, although in 2 dimensions the gains are expected to be less significant due to the reduced amount of work in the counts-in-cells method.

The layout of this paper is as follows: in section , we quickly review the statistics we wish to calculate. This is followed by an explicit description of our new algorithm, and an examination of its performance. Next we present a brief case study on applying our algorithm to cosmology and conclude with a brief summary.

2 Statistics: Moments and Correlation Functions

Due to space limitations a full discussion of the counts-in-cells method, and how it is related to higher order moments, is beyond the scope of this paper. However an excellent discussion of counts-in-cells and statistical measurement processes may be found in Peebles (1980). For completeness, we briefly summarize the statistics we are interested in measuring.

The 2-pt CF, $\xi(r)$, measures the radial excess/deficit over Poisson noise for a point process. It is defined in terms of the joint probability, δP, of finding objects in volume elements δV_1 and δV_2 separated by a radial distance r_{12}, viz,

$$\delta P = n^2(1 + \xi(r_{12}))\delta V_1 \delta V_2, \tag{1}$$

where n is the average number density of the point process. The Fourier transform pair of the 2-pt CF is the power spectrum, $P(|\mathbf{k}|)$,

$$P(|\mathbf{k}|) = \frac{1}{(2\pi)^3} \int_V e^{-i\mathbf{k}.\mathbf{r}} \xi(|\mathbf{r}|) d^3r \tag{2}$$

which is used to describe the statistics of the initial density field in cosmology.

The joint probability idea can be generalized to n-pt processes, for example, the reduced 3-pt CF is defined by;

$$\delta P = n^3 \delta V_1 \delta V_2 \delta V_3 \times$$

$$(1 + \xi(r_1) + \xi(r_2) + \xi(r_3) + \zeta(r_1, r_2, r_3)), \tag{3}$$

where r_1,r_2 and r_3 are defined by the triangle described by the three points under consideration. For cosmology, the assumptions of homogeneity and isotropy require that $\zeta(r_1, r_2, r_3)$ be a symmetric function of these three lengths. Higher order correlation functions follow in a logical manner.

Using the counts-in-cells method, it can be shown that the second central moment $\mu_2 = \left\langle (N - nV)^2 \right\rangle$, where N is the count of points within spheres of radius r (and volume V), is given by

$$\mu_2 = nV + n^2 \int_V dV_1 dV_2 \xi(r_{12}). \tag{4}$$

The third central moment $\mu_3 = \left\langle (N - nV)^3 \right\rangle$, is given by

$$\mu_3 = 3\mu_2 - 2nV + n^3 \int_V dV_1 dV_2 dV_3 \zeta \tag{5}$$

Both these equations show how integrals over the correlation functions enter in to calculations of the central moments. Relationships for the higher order moments can be constructed, but rapidly become lengthy to calculate (*e.g.* Fry and Peebles, 1978).

The final definition we require is one that relates higher order cumulants to the variance. To aid our discussion we introduce the following notation: the over-density of a point process relative to the mean density, $\bar{\rho}$, is given by $\delta(\mathbf{x}) = \Delta\rho(\mathbf{x})/\bar{\rho}$ where $\Delta\rho = \rho(\mathbf{x}) - \bar{\rho}$ is the local deviation from the average density. Although this is most usually recognized as a continuum description, it also provides a useful construct for our discussion of point processes. For example, since the local density of particles in the counts-in-cells method is given by N/V, $\delta(\mathbf{x}) \simeq (N/V - n)/n$. From this definition of $\delta(\mathbf{x})$ the n-th order connected moments of the point process define the 'S_p' statistics via the following definition[1]:

$$\langle \delta^p \rangle = S_p \left\langle \delta^2 \right\rangle^{p-1}. \tag{6}$$

The S_p statistics play a central role in analysis of redshift surveys. To date, up to S_9 has been calculated by researchers (Szapudi et al., 1996).

3 The Smoothed Field Algorithm (SFA)

While the counts-in-cells method is conceptually beautiful in its relation to the S_p statistics, it is computationally strenuous to calculate. As the radius of the sampling sphere becomes larger, on average the work to calculate the count within the sphere will grow at a cubic rate. In reality the situation can be potentially worse, since inefficiencies in particle book-keeping can appear (*i.e.* having to search far down tree-nodes, or equivalently searching through very dense cells in a grid code). To counter this problem one can use a hierarchical (tree) storage of counts in cells on a grid, as discussed in Szapudi et al. (1999). This greatly improves calculation time, since the summation over particles within cells is much reduced at large radii. Using this method it has been reported that 10^9 samples from a data set with 47 million particles can be generated in 8 CPU hours.

The basis of our alternative 'smooth field algorithm' is that each counts-in-cells value is a discrete sample of the local density field smoothed over the scale of the sample sphere. In the continuum limit of an infinite number of particles, defining the density $\delta(\mathbf{x})$, the sampled value $\delta_s(\mathbf{x})$ can be written as an integral over the spherical top-hat function $W(r, r_f)$

$$W(r, r_f) = \begin{cases} 1, & r \leq r_f; \\ 0, & r > r_f, \end{cases} \tag{7}$$

of radius r_f and the raw density field $\delta(\mathbf{x})$, to give,

$$\delta_s(\mathbf{x}) = \frac{1}{V_{TH}} \int_V \delta(\mathbf{x} + \mathbf{r}) W(|\mathbf{r}|, r_f) d^3 r, \tag{8}$$

where V is the volume of the periodic sample region and V_{TH} the volume of the sample sphere (a 3 dimensional top-hat). Via the Convolution Theorem, the Fourier transform of $\delta_s(\mathbf{x})$, namely, $\hat{\delta}_s(\mathbf{k})$ is given by

$$\hat{\delta}_s(\mathbf{k}) = \hat{\delta}(\mathbf{k}) \hat{W}((|\mathbf{k}|, r_f). \tag{9}$$

Thus we can quickly calculate the *entire* $\delta_s(\mathbf{x})$ field by Fourier methods.

The discrete calculation of counts can be expressed in almost the same way, except that the continuous density field is replaced by a discrete sum of three dimensional Dirac delta functions, $\delta^D(\mathbf{x})$,

$$N_s(\mathbf{x}) = \frac{1}{V_{TH}} \int_V \sum_{i=1}^{N_p} \delta^D(\mathbf{x} + \mathbf{r} - \mathbf{x}_i) W(|\mathbf{r}|, r_f) d^3 r \tag{10}$$

where N_p is the number of particles in the simulation, and $\mathbf{x}_i$ gives the position of particle i. In the counts-in-cells method the integral over the volume is replaced by a summation within the given search volume V_{TH}.

To connect these two approaches all that is needed is a smoothing function that will convert a discrete set of points to a continuous density field. We require a smoothing function, $A(\mathbf{x})$, which can be summed over the particle positions to reproduce a smooth field $\delta(\mathbf{x})$. Provided we can do this, we can use Fourier methods to precalculate all of the required $\delta_s(\mathbf{x})$ values and greatly reduce the amount of work. In practice it will be necessary to define a discrete density on a grid, and then use an interpolation process to provide a continuum limit. The smoothing idea has been studied in great depth (see Hockney and Eastwood (1988) for explicit details) and there exists a series of computationally efficient smoothing strategies that have good Fourier space properties, as well as having well defined

[1] The S_p statistics are motivated by the assumption that, given the 2-pt CF, $\xi(r) = (r_0/r)^\gamma$, the n-pt correlation functions scale as $\xi^{(n)}(\lambda x_1, ..., \lambda x_n) = \lambda^{-\gamma(n-1)} \xi^{(n)}(x_1, ..., x_n)$, see Balian and Schaeffer (1989).

interpolation function pairs. The most common smoothing function ('assignment function') mechanisms are 'CIC' (Cloud-in-Cell), and 'TSC' (Triangular Shaped Cloud). Cloud-in-cell interpolation provides a continuous piecewise linear density field, while TSC has a continuous value and first derivative. The only potential issue of difficulty is that sampling a continuous periodic variable at discrete points means that the Fourier domain is finite and periodic and thus has the possibility of being polluted by aliased information (with images separated by $2\pi/L$ where L is the size of the period). In practice, the higher order assignment functions have a sufficiently sharp cut-off in Fourier space that this is not a significant problem[2].

Having established that we can convert our discrete set of points into a continuous density defined by a grid of values and an interpolation function, we must decide upon the size of grid to be used. The initial configuration of points (corresponding to a low amplitude power spectrum) is such that the majority of neighbouring particles have separations close to the mean inter-particle separation $N_p^{1/3}$. Therefore, for this configuration we use a grid defined such that $L^3 = N_p$. This is beneficial on two counts: firstly, the grid requires a comparatively small amount of memory to store than the particle data, and secondly, it captures almost all the density information stored in the particle distribution (since most particles are separated by sizes close to the grid spacing).

To summarize, the steps in the SFA are as follows:

1. Use an assignment function, $A(\mathbf{x})$, to smooth the mass (m) associated with each of the particles on to a grid. This creates the grid representation of the density field, $\rho(\mathbf{x})$:

$$\rho(\mathbf{x}) = \frac{m}{V}\sum_{i=1}^{N_p} A(\mathbf{x}_i - \mathbf{x})$$

2. Fourier transform the density field $\rho(\mathbf{x})$ to form $\hat{\delta}(\mathbf{k})$

3. Multiply by $G(k)$, the product of the Fourier transform of the real space top-hat filter ($\hat{W}(k, r_f) = 3(\sin(kr_f) - kr_f\cos(kr_f))/kr_f^3$) and the inverse of the assignment function filter, which includes an alias sum out to two images

4. Fourier transform the resulting field back to real space

5. Calculate $\delta_S(\mathbf{x})$ at all sampling positions using the interpolation function pair to the original assignment function $A(\mathbf{x})$

6. Calculate desired statistics

In this paper we have used a 3rd order polynomial assignment function ('PQS', see Hockney and Eastwood, 1988) which is defined (in 1-dimension) by;

$$A_1(x) = \begin{cases} \frac{2}{3} + |x|^2\left(\frac{|x|}{2} - 1\right), & |x| \le 1; \\ \frac{1}{6}(2 - |x|)^3, & 1 < |x| \le 2; \\ 0, & |x| > 2, \end{cases} \qquad (11)$$

and the 3-dimensional function is defined $A(x, y, z) = A_1(x)A_1(y)A_1(z)$. Note that $A(x, y, z)$ is not an isotropic function, which in this case is beneficial for speed, since it is unnecessary to calculate a square root. It also simplifies calculating the Fourier transform of the assignment function since all the dimensions are now separable. Note that $A_1(x)$ has a comparatively wide smoothing profile, and therefore its Fourier transform is a strongly peaked function with good band-limiting properties. This is advantageous for dealing with the aliasing problem mentioned earlier. Indeed, the Fourier transform of $A_1(x)$ is:

$$\hat{A}(k_x) = \left(\frac{\sin(k_x/2)}{k_x/2}\right)^4, \qquad (12)$$

which has a $1/k^4$ suppression of power. This is sufficiently sharp to ensure that only the first and second images need be accounted for in $G(k)$ (the Green's function associated with top-hat filtering and the assignment process).

4 Performance Comparison

Before proceeding to parallelize the algorithm, it is instructive to compare the speed of the serial algorithm as compared to the counts-in-cells method. In figure 2 we show the time to calculate 2.1×10^6 samples on 2.6×10^5 points as a function of the sample radius. A (logarithmic) least-squares fit showed that the time for the standard counts-in-cells method (version 1) grows as $r^{2.5}$, which is slightly lower than the expected value of r^3. For the second counts-in-cells algorithm we developed, which is optimized by storing a list of counts in the chaining cells used to control particle book-keeping in the code, the dependence with radius was found to be r^2. This is understood from the perspective that most of the work in each sample has already been performed in the summation within chaining cells and that the work for each sample thus becomes dependent on sorting over the cells at the surface of the sample area, which is proportional to r^2. However comparison of both these methods to the SFA shows they are far slower in comparison. Because the entire $\delta_s(\mathbf{x})$ field is precalculated (modulo the interpolation process to non-grid positions) in the SFA method, the time to calculate the samples is constant as a function of radius, and is exceptionally fast. Based up the data presented in figure 2, we initially estimated being able to calculate 10^9 sample points on a 512^3 data set in less than 2 CPU hours, which is over 4x faster than the results reported for tree-optimized counts-in-cells methods (Szapudi et al., 1999). We have recently confirmed this result using our parallel code, which took 6.5 minutes on

[2]See Hockney and Eastwood (1988) for a discussion of this point. Aliases can only be removed completely by assigning information to all points on the sampling grid for each point/particle, which it too computationally expensive to be feasible.

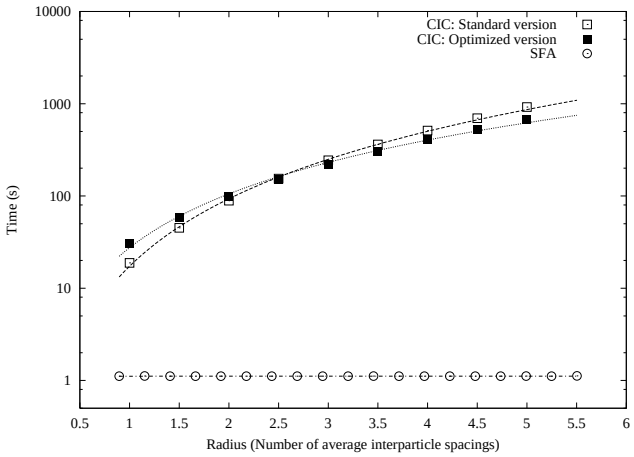

Figure 2: Comparison of speed for two versions of the counts-in-cells (CIC) method versus SFA for 2.6×10^5 particles and 2.1×10^6 sample points at different smoothing radii. Least squares fits are given for all data. The first counts-in-cells method is a straight summation over particles contained within the sampling sphere, while the second method is optimized to store a count of particles within chaining cells. Provided a chaining cell lies within the radius of the sampling sphere then the sum within the chaining cell is not necessary.

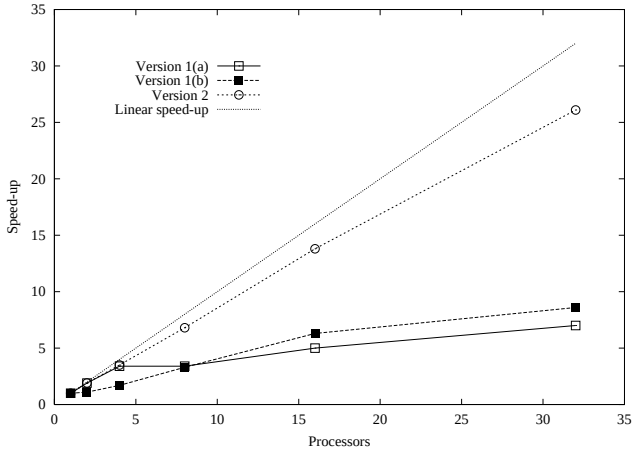

Figure 3: Comparison of speed-ups for different implementations of the SFA. Version 1(a) is the standard method with block data decomposition but no directed assignment within RADs. Version 1(b) forces data and threads to spread across RADs. Version 2 corresponds to our data local sampling and is clearly superior.

32 processors to calculate 10^9 samples on a 512^3 particle data set produced for a project being conducted at the Pittsburgh Supercomputing Center.

5 Parallelization

Typically when calculating statistics, the value of the sampling radius (equivalently the top-hat radius) is varied so that the entire sampling process must be repeated many times. Thus the most obvious method of parallelization is to create several different grids for each smoothing radius and process them in parallel. However, available memory considerations may well make this impractical. Instead, it is better to parallelize each calculation for each radius. This is non-trivial as the following algorithmic steps must be parallelized:

1. Calculation of Green's function
2. Forward FFT of density grid to k-space
3. Multiplication of density grid by Green's function
4. Reverse FFT to real space
5. Sum over sample points

The first four items have all been parallelized previously for our main simulation code (see Thacker et al., 1998). The final step, while appearing to be somewhat straightforward, must be approached with care (as we shall demonstrate). The obvious issues which need to addressed are (1) ensuring each thread has a different random seed for sample positions and (2) that the sum reduction of the final values across threads is performed. In practice, both of these issues can be dealt with in very straightforward ways using the OpenMP shared memory programming standard. Sum reductions can be controlled via the `REDUCTION` primitive while different random seeds can be set using an array of initial values. Parallelization in this environment turned out to be straightforward.

Tests on a 32 processor HP GS320 (1 GHz Alpha EV6/7 processors) at the Canadian Institute for Theoretical Astrophysics (CITA), showed reasonable speed-up (see figure 3), but comparatively poor efficiency (22%) when 32 processors were used. There is also a noticeable step in the speed-up at 4 to 8 processors. This step is caused by memory for a job being moved to a second memory domain, or 'resource affinity domain' (RAD), within the machine. The 32 processor machine has 8 RADs in total, connected via a cross-bar, with 4 processors belonging to each RAD. Latency to remote RADs is significantly higher than to local RADs, which explains the increased execution time. Additionally, as the amount of traffic on the cross-bar between the RADs increases, latencies are known to increase by very large factors (up to 3000 nanoseconds, Cvetanovic, 2003). This is a serious bottleneck in the GS320 design which has been removed in the latest GS1280 machine. Ultimately, to improve performance on the GS320, it is necessary to increase the locality of the sampling technique to reflect the locality of memory within the machine, and avoid sending data across the cross-bar.

Note that using a block decomposition of data across the RADs means that locality is only really necessary in one axis direction. Therefore, we adopted the following

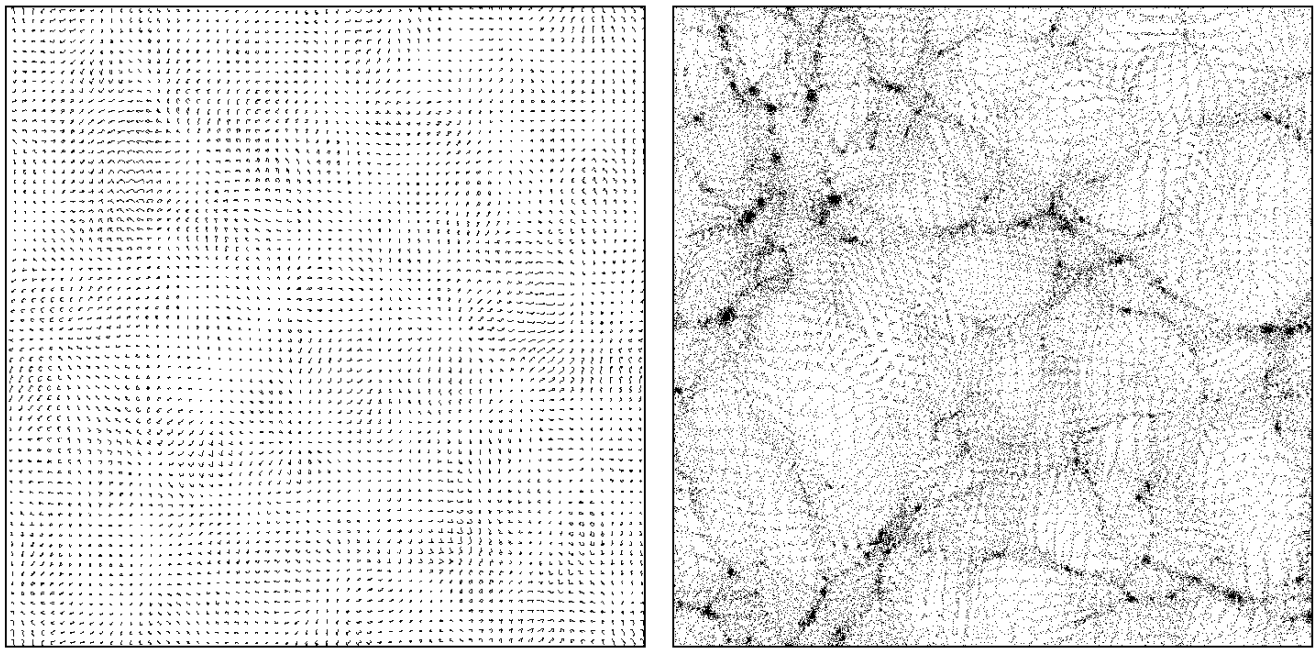

Figure 4: Initial and final point configurations for a slice through a simulation with 2.6×10^5 particles.

strategy to improve performance:

1. Block decomposition of the $\delta_s(\mathbf{x})$ grid across RADs
2. Pre-calculate the list of random positions in the z-axis
3. Parallel sort the list of random positions in increasing z value
4. Parallelize over the list of z positions, calculating x and y values randomly

The resulting sample still exhibits Poisson noise statistics and is therefore valid for our purposes. However, the sample points are now local in the z direction, which greatly reduces the possibility of remote access due to the block assignment of data. The scaling improvement for this method is shown in figure 3. The improvement is striking. We achieved a 1.2x increase in performance for the single processor result alone, while at 32 processors we have achieved a 4.8× improvement in speed-up and a tripling of the parallel efficiency (82%). Note that the speed-up is still not perfect for the improved version. This may be a bandwidth issue since the interpolation at each sampling point requires 64 grid values, which breaks down into 16 cache lines, with only 8 floating point calculations performed for all the data in each cache line. Note that it is unlikely that using the next lowest level of interpolation (TSC) would help. TSC requires 27 points grid points per sample, which is 9 cache lines, with 6 floating point calculations per cache-line. Thus the overall ratio of calculation to memory fetches is actually reduced.

6 Application: Moments in Initial Conditions

The initial conditions for cosmological structure are prescribed by initial density, temperature and velocity fields. Although there is debate over whether evolution in the early universe (such as magnetic fields) may induce a non-Gaussian signal in the initial conditions (White et al., 1994), most researchers believe that the density field is Gaussian process, and the velocity may be derived directly from it. In the absence of non-Gaussian features, the density field, which is usually discussed in terms of the linear over-density δ, is completely described by its continuous power spectrum $P(k) = Ak^n$, where A is a normalization constant. This initially smooth field evolves under gravity to produce the locally inhomogeneous and biased distribution of galaxies we observe today (see figure 4, which compare particles positions from initial to final outputs). Early evolution, when $\delta(\mathbf{x}) \ll 1$, is in the linear regime and can be described by perturbation theory. As the over-density values approach and later exceed unity, it is necessary to use simulations to calculate the non-linear evolution. Thus, ideally, the initial conditions for simulations should correspond to the latest time that can be followed accurately by perturbation theory.

Scoccimarro (1998) has developed an algorithm for the fast calculation of the particle positions required for cosmological simulations via 2nd order Lagrangian perturbation theory (2LPT). We have recently implemented this algorithm in parallel using OpenMP. Although 2LPT requires more computation, it has significant advantages over the standard 1st order technique (known as the Zel'dovich (1968) approximation) as higher order moments exhibit far less transient deviations at the beginning of the simulation. Further, one should in principle be able to follow the initial evolution to slightly later epochs using 2LPT and therefore

begin simulations at a slightly later time. In practice, the transient deviation issue is most significant.

In general, the more negative the spectral index the faster the initial transients die away. This is helpful, since most simulations are conducted with an effective spectral index, n, of between -1.5 to -3 (depending on the size of the simulation volume). Also, although we have focused solely on particle position statistics in this paper, it is worth noting that a similar analysis can be applied to velocity fields defined on the point process. Analysis of the transients in the velocity divergence field, $\theta = \nabla.v$, shows an even greater improvement when using the 2LPT method (Scoccimarro, 1998).

To test whether our new 2LPT code was reproducing the correct results we have compared the measured S_3 statistics for our 2LPT initial conditions versus those produces with the Zel'dovich approximation (1st order). At the initial expansion factor of $a = 1$, the ZA predicts the following value for S_3 (Bernardeau, 1994);

$$S_3 = \frac{28}{7} - (3 + n), \tag{13}$$

while 2LPT predicts;

$$S_3 = \frac{34}{7} - (3 + n). \tag{14}$$

Thus after performing the 2nd order correction the value of S_3 should increase by 6/7. In figure 5 we show the calculated values of S_3 for two sets of initial conditions, one created using the ZA and the other with the additional 2LPT correction. Both the SFA measured values of S_3 are high for this particular set of phases (as compared to the theoretical prediction), but we have confirmed that alternative random seeds can produce similar results. Indeed we have found the values of S_3 are quite dependent upon the phases of the Fourier waves used, and achieving a value that is asymptotic to the theoretical value is extremely difficult. We are currently investigating this phenomenon in more detail. However, a brief visual inspection of figure 5 provides evidence that the residual, Δ, between the ZA and 2LPT results is close to $6/7 \simeq 0.86$. Analysis of the set of residuals between the two lines gives $\Delta = 0.88 \pm 0.03$ (1σ deviation), confirming that our code is accurately reproducing the difference in S_3 values.

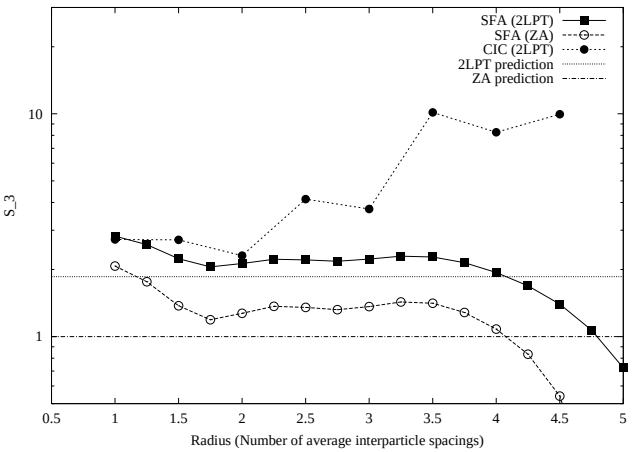

Figure 5: Comparison of S_3 calculated via CIC versus SFA on an n=0 initial condition, with the theoretical result shown for reference. SFA shows a good match out to 4 inter-particle spacings at which point it begins declining. CIC appears accurate on small scales but rapidly diverges away from the true signal. We have confirmed that as the simulation evolves, and the effect of shot noise is reduced, both methods converge to similar values.

7 Summary and Discussion

We have presented a new fast algorithm for rapid calculation of one point cumulants for point processes. Our algorithm is based upon a smoothed field approach, which reproduces the underlying statistical properties of the point processes field from which it is derived. The method is significantly faster than counts-in-cells methods because the overhead of evaluating the number of particles in a given sphere has been removed. We are able to calculate 10^9 sample points on a 512^3 data set in less than 2 CPU hours, which is over 4x faster than the results reported for tree-optimized counts-in-cells methods (Szapudi et al., 1999). We also note that while tree methods also lead to very large speed ups, they are still subject to noise from the point process for low amplitude signals.

We are currently applying this new technique to examine the evolution of high order moments in cosmological density fields at low amplitude levels and will present our findings elsewhere (Thacker, Couchman and Scoccimarro in prep). We also anticipate making the codes described in this paper publically available in the near future.

8 Acknowledgments

RJT is partially supported by a CITA National Fellowship. HMPC acknowledges the support of NSERC and the CIAR. RJT would like to thank Evan Scannapieco and Lars Bildsten for hosting him at U. C. Santa Barbara where part of this research was conducted. This research utilized CITA and SHARCNET computing facilities.

REFERENCES

Peacock, J.A. (1998) *Cosmological Physics*, Cambridge University Press.

Pais, A. (1982) *Subtle is the Lord: The Science and the Life of Albert Einstein*, Oxford University Press.

Bennett, C.L. *et al.* (2003) 'First-Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Preliminary Maps and Basic Results', *The Astrophysical Journal: Supplements*, Vol. 148, No. 1, pp.1–27.

Kolb, E.W. and Turner, M.S. (1990) *The Early Universe*, Addison Wesley.

Sanders, R.H. and McGaugh, S.S. (2002) 'Modified Newtonian Dynamics as an Alternative to Dark Matter', *Annual Reviews of Astronomy and Astrophysics*, Vol. 40, pp.263–317.

Thacker, R.J. and Couchman, H.M.P (2001) 'Star Formation, Supernova Feedback, and the Angular Momentum Problem in Numerical Cold Dark Matter Cosmogony: Halfway There?', *The Astrophysical Journal*, Vol. 555, No. 1, pp.L17-L20.

Hubble, E.P. (1929) 'A Relation between Distance and Radial Velocity among Extra-Galactic Nebulae', *Proceedings of the National Academy of Science*, Vol. 15, No. 3, pp.168–173.

2-degree Field Galaxy Survey. [Online 2004]. Available: http://www.mso.anu.edu.au/2dFGRS/

Sloan Digital Sky Survey. [Online 2004]. Available: http://www.sdss.org/

Peebles, P.J.E (1980) *The Large Scale Structure of the Universe* Princeton University Press.

Couchman, H.M.P. and Peebles, P.J.E. (1998) 'A Numerical Renormalization Solution for Self-similar Cosmic Structure Formation', *The Astrophysical Journal*, Vol. 497, No. 2, pp.499–511.

Thacker, R.J., Pringle, G., Couchman, H.M.P. and Booth, S. 'HYDRA-MPI: An Adaptive Particle-Particle, Particle-Mesh code for conducting Cosmological Simulations on Massively Parallel Architectures', *High Performance Computing Systems and Applications 2003* NRC Research Press.

Fry, J.N. and Peebles, P.J.E. (1978) 'Statistical analysis of catalogs of extragalactic objects. IX - The four-point galaxy correlation function', *The Astrophysical Journal*, Vol. 221, No. 1, pp.19–33.

Balian, R. and Schaeffer, R. (1989) 'Scale-invariant matter distribution in the universe. I - Counts in cells', *Astronomy & Astrophysics*, Vol. 220, No. 1, pp.1–29.

Szapudi, I., Meiksin A. and Nichol, R.C. (1996) 'Higher Order Statistics from the Edinburgh/Durham Southern Galaxy Catalogue Survey. I. Counts in Cells', *The Astrophysical Journal*, Vol. 473, No. 2, pp.15–21.

Szapudi, I., Quinn, T., Stadel, J. and Lake, G. (1999) 'N-Point Correlations in CDM and OmegaCDM Simulations', *The Astrophysical Journal*, Vol. 517, No. 1, pp.54–63.

Zel'dovich, Ya.B., (1970) 'Gravitational instability: an approximate theory for large density perturbations', *Astronomy & Astrophysics*, Vol. 5, pp.84–89.

Scoccimarro, R. (1998) 'Transients from initial conditions: a perturbative analysis',*Monthly Notices of the Royal Astronomical Society*, Vol. 299, No. 4, pp.1097–1118.

Thacker, R.J., Couchman, H.M.P. and Pearce, F.R. (1998) 'Simulating Galaxy Formation on SMPs', *High Performance Computing Systems and Applications 1998*, Kluwer Academic.

Hockney, R.W. and Eastwood, J.W. (1988) *Computer Simulation Using Particles*, McGraw-Hill.

Cvetanovic, Z. (2003) Performance Analysis of the Alphaserver GS1280. [Online 2003]. Available: http://www.uni-magdeburg.de/urzs/marvel/marvel_performance.pdf

White, M., Scott, D. and Silk, J. (1994) 'Anisotropies in the Cosmic Microwave Background',*Annual Reviews of Astronomy and Astrophysics*, Vol. 32, pp.319–370.

Bernardeau, F. (1994) 'Skewness and kurtosis in large-scale cosmic fields', *The Astrophysical Journal*, Vol. 433, No. 1, pp.1–18.